Prise en compte de l'importance d'un site web dans l'estimation de la probabilité a priori de pertinence d'une page web

Arezki Hammache[(1)], Mohand Boughanem[(2)], Rachid Ahmed-Ouamer[(1)]

[(1)] Laboratoire LARI, Département d'informatique, Université Mouloud Mammeri
15000 Tizi-Ouzou, Algérie
{arezki20002002,ahm_r}@yahoo.fr
[(2)] Laboratoire IRIT, Université Paul Sabatier
118 route de Narbonne 31062 Toulouse Cedex 09, France.
bougha@irit.fr

RÉSUMÉ. *Plusieurs caractéristiques ont été utilisées pour estimer la probabilité a priori d'un document comme : la longueur du document, la structure des liens, le facteur temps. Cependant, ces caractéristiques dépendent seulement du document lui même. Or, dans le contexte du web une page web fait partie en général d'un site web. L'idée que nous explorons dans cette article est l'utilisation des caractéristiques du site contenant la page concernée pour conditionner la probabilité a priori de pertinence de la page. Une fois cette probabilité est calculée nous la combinons avec le score obtenu par le contenu de la page web. Cette combinaison des deux évidences est réalisée sous le cadre de modèle de langue. Afin de valider notre idée, nous avons effectué des tests sur la collection TREC « .GOV » ; où nous avons comparé les différentes versions de notre modèle avec deux modèles : le modèle uni-gramme qui ne considère que le contenu de la page, et le modèle combinant le contenu d'une page web et la probabilité a priori de la page obtenu en utilisant seulement une caractéristique sur la page (nombre de liens entrants). Les résultats obtenus montrent que notre modèle est très prometteur.*

ABSTRACT. *Several features have been used to estimate the priori probability of a document such as: the document length, the link structure, the time factor. But, these features depend only on the document. However, in the context of web, a web page is part of website. The idea that we explore in this article is to use website features which contains the concerned page to estimate the page prior. Once this probability is calculated we combine it with the score obtained by the content of the web page. This combination is performed under the language model framework. To validate our idea, we carried out tests on the ". GOV" TREC collection where we compared different versions of our model with two models: the uni-gram model, which considers only the content of the page, and the model combining the content of a web page and the prior probability of the page obtained using only the document feature (the number of in link) . The results obtained show that our model is very promising.*

MOTS-CLÉS : *Recherche d'Information sur le Web, Modèle de Langue, Probabilité a priori de Pertinence.*

KEYWORDS: *Web Information Retrieval, Language Model, Document Prior*

1. Introduction

La plupart des moteurs de recherche d'information sur le web privilégient la minimisation du temps de réponse par rapport à la qualité des documents retournés à l'utilisateur. En effet, ceux-ci délivrent des résultats massifs en réponse aux requêtes des utilisateurs, ce qui génère ainsi, une surcharge informationnelle dans laquelle il est difficile de distinguer l'information pertinente de l'information secondaire ou même du bruit. L'une des raisons qui a engendré ceci est du à la non prise en compte de toutes les dimensions d'un document web dans le processus d'indexation et de recherche. En effet, les moteurs de recherche implémentent les techniques traditionnelles de la RI. Cependant, les documents web, généralement sous format Html, sont des documents structurés via des balises et interconnectés par des liens hypertextes, de plus un document web traite un ou plusieurs thèmes exprimés implicitement par les liens entre les termes du document.

Dans cet article nous traitons de l'exploitation de la structure des hyperliens pour estimer la probabilité a priori de pertinence d'un document web. En se basant sur le postulat qu'une page web plus populaire est plus probable d'être pertinente qu'une page moins populaire.

Le modèle de langue est un nouveau cadre probabiliste pour la description du processus de la recherche d'information (RI). Le modèle de langue convient bien à cette nouvelle dimension de la RI (i.e. : la RI sur le web), du fait que ce modèle se base sur des fondements mathématiques (statistique et probabilité) qui permettent de mieux analyser et modéliser cette énorme masse de documents (le web). De plus ce modèle offre un cadre qui permet de combiner différentes représentations d'un document. Entre autres le contenu du document et les informations a priori de la pertinence d'un document.

Plusieurs caractéristiques des documents web ont été explorées pour estimer la probabilité a priori de pertinence d'un document, comme : la longueur de document, la structures des liens, etc. Nous proposons dans cet article l'utilisation de la structure du web pour estimer cette probabilité a priori, en se basant sur la définition basique du web qui est considéré comme un ensemble de sites web, lesquels sont composés d'un ensemble de pages web.

Nous organisons ce papier comme suit : La section 2 traite l'état de l'art, deux points essentiels sont abordés : la modélisation de langue en recherche d'information et l'intégration des informations sur la pertinence a priori du document. La section 3 est consacrée à la présentation de l'approche que nous adoptons pour estimer et intégrer les informations a priori de pertinence d'un document web dans le modèle de langue. Nous rapportons les résultats expérimentaux dans la section 4. Enfin, dans la section 5, nous concluons notre travail et énumérons quelques perspectives.

2. Etat de l'art

2.1 Le modèle de langue

Le modèle de langue est un cadre probabiliste pour la description du processus de la RI (Ponte *et al.*, 1998). Les résultats obtenus avec ce modèle ont montré des performances équivalentes voire supérieures à celles des modèles classiques (vectoriel, probabiliste) (Zhai *et al.*, 2004). L'estimation de score d'un document d vis-à-vis d'une requête q est réalisée dans le modèle de langue comme suit :

$$P(q|d) = P(d)\prod_{t_i \in Q} P(t_i|d) \quad (1)$$

L'évaluation de modèle de document $P(t_i|d)$ peut être réalisée en utilisant n'importe quel modèle de langue uni-gramme. Dans ce travail le lissage Dirichlet est utilisé. Le modèle est exprimé ainsi:

$$P_{Dir}(t_i|d) = \frac{F(t_i, d) + \mu P(t_i|C)}{|d| + \mu}$$

Où $F(t_i, d)$ est la fréquence de terme dans le document d, $P(t_i|C)$ est le modèle de langue de la collection, $|d|$ est la longueur du document et μ est un paramètre de lissage.

Le facteur $P(d)$ de la formule (1) représente la probabilité a priori de pertinence d'un document. Si aucune caractéristique sur le document n'est utilisée pour l'estimer, alors elle peut être ignorée lors du classement des documents. Par contre si une caractéristique sur un document est utilisée pour son estimation alors les documents de la collection n'ont pas la même probabilité a priori, par conséquent le classement des documents est affecté par l'introduction de ce nouveau facteur.

2.2 La pertinence a priori d'un document

Selon l'approche adoptée, les propriétés (taille de document, nombre de liens entrants, etc.) indépendantes des requêtes peuvent être utilisées ou pas pour conditionner la probabilité a priori de pertinence d'un document. Si la probabilité a priori de pertinence $P(d)$ n'est pas conditionnée par l'une de ces propriétés alors cette probabilité représente la probabilité de prélever un document de la collection. Par conséquent tous les documents dans la collection ont la même probabilité d'être sélectionnés, et donc la probabilité a priori de pertinence de document peut être ignorée lors du classement des documents. Par contre si la probabilité a priori est conditionnée par l'une de ces caractéristiques alors les documents de la collection n'ont pas la même probabilité a priori, et donc les documents ne sont pas équiprobables. Par exemple, si la caractéristique utilisée est le score de popularité de document alors un document populaire est plus probable d'être pertinent qu'un document moins populaire.

Plusieurs caractéristiques ont étés utilisées pour estimer la probabilité a priori d'un document, on peut citer : la longueur du document (Kraaij *et al*., 2001) (Parapar *et al*., 2009) la structures des liens (Kraaij *et al*., 2001) (Hauff *et al*., 2005), le facteur temps (Li *et al*., 2003) (Diaz *et al*., 2004) et le rapport information/bruit (Zhu *et al*., 2000). L'intuition derrière l'utilisation de ces caractéristiques est que : un document est plus probable d'être pertinent car : il est plus long, il est plus populaire, il est plus récent, ou contient plus d'informations que de bruit. Nous présentons ci-dessous quelques travaux utilisant ces caractéristiques.

La structure des liens : L'intuition derrière l'utilisation de la structure des liens est que les documents populaires ou les plus cités tendent à être plus pertinents. La méthode simple d'utilisation de la structure des liens est l'usage du nombre de liens entrants. La probabilité de pertinence a priori est exprimée comme suit :

$$P(d) = \frac{n(l,d)}{\sum_{d_i} n(l,d_i)}$$

Où $n(l,d)$ est le nombre de liens entrants dans le document d.

D'autres facteurs plus sophistiqués on été utilisés comme : le PHits (Cohn, *et al*, 2000), le PageRank (Brin *et al*, 1998) qui est à l'origine du moteur de recherche Google. Le principe de cet algorithme consiste à ordonner les pages web selon leur popularité, en se basant sur l'hypothèse suivante : « une page est populaire (importante) quand elle est beaucoup citée ou citée par une page très populaire». L'estimation de cette popularité est formalisée comme suit :

$$PR(p) = (1-d) \times \frac{1}{T} + d \times \sum_{i=1}^{k} \frac{PR(p_i)}{C(p_i)}$$

Où :

$PR(p)$: est le PageRank de la page p ;

T : est le nombre total de pages sur le web (indexées) ;

d : est un paramètre fixé à 0.85 ;

$C(p_i)$: est le nombre de liens sortant de la page p_i, et k est le nombre de pages qui pointent la page p.

De nombreux travaux ont montré que l'incorporation du PageRank dans le classement des pages web, améliore les performances de recherche sur de grandes collections de type web (Upstill *et al*, 2003) (Craswell *et al*, 2005) (Peng *et al*, 2007). Cependant, ces algorithmes (PageRank et Hits) ne considèrent pas l'aspect hiérarchique du web.

La taille du document : La probabilité de pertinence a priori est proportionnelle à la taille du document, elle est exprimée ainsi:

$$p(d) = \frac{|d|}{|C|}$$

Où $|d|$ est la taille de document et $|C|$ est la taille de la collection.

L'intuition de l'utilisation d'une telle caractéristique est qu'un document plus long tend à contenir plus d'informations et par conséquent il est plus probable d'être pertinent. Les résultats obtenus avec l'utilisation de cette caractéristique ont été mixtes et cela selon la collection utilisée (Kraaij *et al*., 2001).

Parapar et al (Parapar *et al*., 2009) ont proposé d'estimer cette probabilité $P(d)$ en utilisant la taille compressée d'un document. La formule utilisée est exprimée ainsi :

$$P(d) = \frac{comp(d)}{\sum_{d_i \in C} comp(d_i)}$$

Où $comp(d)$ est la taille en octets du document d compressé (zippé) divisée sur la taille originale en octets du document d. Ce nouveau facteur a été évalué et comparé au facteur taille originale de document en utilisant quatre collections TREC. Les résultats présentés montrent que la taille compressée d'un document obtient des améliorations de précision moyenne (MAP) allant de +0,4% à +3,1% par rapport à l'utilisation de la taille originale de document.

La date de création du document : D'autres travaux utilisent l'intuition suivante : « Les documents récents tendent à être plus pertinents que les documents anciens » pour estimer la probabilité a priori d'un document, Li et Croft (Li *et al*., 2003) ont proposé un modèle de langue qui permet d'intégrer la notion de « temps » dans l'évaluation de pertinence d'un document vis-à vis d'une requête, où ils assignent une plus grande probabilité de pertinence pour les documents ayant une date de création récente. Ainsi, ils expriment La probabilité de pertinence a priori d'un document sachant sa date de création, comme une distribution exponentielle, exprimée ainsi:

$$P(d|T_d) = \lambda e^{-\lambda(T_C - T_d)}$$

Où T_C est la date la plus récente dans toute la collection (exprimée en mois) et T_d est la date de création de document d.

Les évaluations réalisées sur un ensemble particulier de requêtes montrent que l'incorporation de la notion de temps en utilisant la distribution exponentielle est bénéfique pour la RI.

Le rapport information/bruit : Il est défini comme le rapport entre la taille de document après prétraitement (élimination des mots vides et des balises HTML) et la taille de document sans prétraitement (Zhu *et al*., 2000). La probabilité de pertinence a priori est exprimée ainsi :

$$P(d) = \frac{l_{token}}{l_{document}}$$

Où l_{token} est la taille de document après le prétraitement et $l_{document}$ est la taille de document avant le prétraitement. Ainsi, un document avec moins de mots vides et peu de balises HTML produit un haut rapport information/ bruit, ce qui signifie que le document est de « bonne » qualité. Une amélioration de précision de l'ordre de + 14,7% est obtenue dans le contexte de la recherche centralisée.

Type d'URL du document : Kraaij et al (Kraaij *et al*., 2001) ont utilisé la forme (type) de l'URL pour estimer la probabilité qu'une page soit une page d'entrée. Elle est définie ainsi :

$$P(d) = P\left(PE|url_{type}(d)\right) = \frac{c(PE, t_i)}{c(t_i)}$$

Où url_{type} est le type de l'URL de document d, $c(PE, t_i)$ est le nombre de documents du type d'URL « t_i» qui sont des pages d'entrées « PE » pour un site web, il est obtenu à partir des évaluations de pertinence et $c(t_i)$ est le nombre de documents de type d'URL « t_i». Quatre types de catégories d'URL ont été définis : Racine, sous-racine, chemin (répertoire) et Fichier. Sur la base de ces quatre types d'URL, ils ont mené des expérimentations sur la collection web WT10g (collection utilisée dans TREC 2001) pour estimer la probabilité qu'une page soit une page d'entrée sachant son type d'URL. Ils ont constaté que cette source d'information est un bon indicateur pour prévoir la pertinence d'une page.

3. Approche proposée

Les caractéristiques utilisées jusqu'ici pour estimer la probabilité a priori $P(d)$ d'un document dépendent du document web (page) uniquement. Or, dans le contexte du web, un document (page) fait partie en général d'un site web lequel fait partie du web. L'idée que nous explorons dans cet article est l'utilisation des caractéristiques du site web qui contient la page concernée pour conditionner la probabilité a priori de pertinence de la page. On s'est basé sur l'hypothèse suivante : « dans la plupart des cas les auteurs des pages web référencent la page principale (site) au lieu de référencer la page exacte (la page concernée) », donc l'utilisation du nombre de liens entrants ou de facteurs plus sophistiqués (comme le PageRank) ne reflète pas l'importance de la page web dans l'espace web. Autrement dit on doit assigner plus de confiance aux pages provenant de sites importants « intéressants » que celles provenant de sites non importants ou même des sites spam.

En partant de l'intuition qu'un site important (référencé par beaucoup d'autres pages) procure une information plus pertinente qu'un site moins important, nous

introduisons le facteur importance du site dans le calcul de l'importance de la page web. Avant de décrire notre modèle, nous définissons ci-dessous les termes clés utilisés.

Un site web (page d'entrée) : est une page dont l'URL contient uniquement, soit le nom de domaine ou un de sous domaine. Par exemple :

www.nist.gov/ et expect.nist.gov/ : sont deux URL de site web.

Une page web : est une page dont l'URL contient un nom de domaine ou un nom de domaine ou sous domaine suivi d'un ou plusieurs répertoires et se terminant (ou non) d'un nom de fichier. Par exemple, les deux sites web (URL) illustrés dans l'exemple précédent contiennent respectivement les pages suivantes :

www.nist.gov/srd/jpcrd_28.htm et expect.nist.gov/scripts/faxstat

Importance d'une page (site) web : l'importance d'un page web est mesurée par le nombre de liens pointant cette page. Tous les liens sont pris en compte, les liens inter-site et liens intra-site.

3.1 Première Version

Nous partons du postulat suivant : « les pages d'un site web permettent de détailler le contenu de la page d'accueil (site) ». Nous proposons alors dans cette version de notre modèle que l'importance d'un site web soit héritée équitablement par l'ensemble des pages web de ce site. De ce fait on estime la probabilité a priori de pertinence d'une page web $P(d)$ ainsi :

$$P(d) = C[\lambda((NB\_s)/N) + (1-\lambda) \times NB\_p] \quad (2)$$

Où : NB_s est le nombre de liens pointant le site « page principale », NB_p est le nombre de liens pointant la page concernée (p), N est le nombre de pages dans le site contenant la page (p) ; C est une constante et λ est un poids compris entre 0 et 1.

La table ci-dessous montre un exemple illustratif de calcul de la probabilité a priori de pertinence suivant la première version de notre approche

Pages	NB_s	NB_p	N	λ	P(d)	Rang1	Rang2
P1	5	5	10	0.5	2.75	2	3
P2	5	2	10	0.5	1.25	3	4
P3	40	6	10	0.5	5	1	2
P4	200	1	20	0.5	5.5	4	1

Table 1. Exemple illustratif

Où : Rang1 est le rang de la page en considérant uniquement son importance (NB_p). Rang2 est le rang de la page en utilisant la formule (2). Nous pouvons remarquer que le rang des pages change ; par exemple la page P4 classée 4ème a passé à la 1ère place car elle appartient à un site important (200 liens entrants).

3.2 Seconde Version

Le deuxième cas que nous explorons est basé sur l'hypothèse que l'information dans la page d'accueil (site) est souvent détaillée dans les pages descendantes directes. Ainsi on réécrit la formule (2) comme suit. Celle-ci est appliquée pour estimer la probabilité a priori de pertinence des pages se trouvant au niveau un (descendantes directes du site):

$$P(d) = C[\lambda((NB\_s)/N_p) + (1-\lambda) \times NB\_p] \quad (3)$$

Où : N_p est le nombre de pages descendantes directes du site. Les autres paramètres sont identiques à ceux de la formule (2).

3.3 Troisième version

En partant de l'hypothèse que deux pages ayant un contenu sémantique similaire devraient avoir une pertinence assez proche pour un même sujet. Alors le fait qu'une page ait un contenu informationnel similaire à celui de son site lui permet de bénéficier de l'importance de site plus que les autres pages. Ainsi, nous formulons la probabilité a priori de pertinence d'une page comme suit :

$$P(d) = C[\left(\frac{sim(d,S)}{\sum_{d_i \in S} sim(d_i,S)}\right) \times NB\_s + (1 - \frac{sim(d,S)}{\sum_{d_i \in S} sim(d_i,S)}) \times NB\_p] \quad (4)$$

Où : $sim(P_i, S)$ représente la similarité sémantique entre la page P_i et le site S auquel elle appartient.

Remarque : pour le calcul de la similarité entre une page et son site, nous utilisons la mesure de cosinus, exprimée comme suit:

$$Sim(P_i, S) = \frac{S.P_i}{\|S\|.\|P_i\|} = \frac{\sum_{j=1}^{|T|} w_{Sj}.w_{P_ij}}{\sqrt{\sum_{j=1}^{|T|} {w_{Sj}}^2 \times \sum_{j=1}^{|T|} {w_{pij}}^2}}$$

Où :

w_{Sj} : est le poids du terme t_j dans le document (site web S).

w_{P_ij}: est le poids du terme t_j dans le document (page web P_i).

$|T|$: est le nombre de termes dans la collection.

4. Résultats expérimentaux

4.1 Collections de test et configuration expérimentale

L'évaluation de notre modèle décrit dans la section 3 précédente est réalisée en utilisant la collection de test TREC .GOV. La Table 2 ci-dessous montre quelques statistiques sur la collection et les requêtes utilisées. Seule la partie titre des requêtes est utilisée.

Collection	#Documents	Requêtes
.GOV	1247753	1-225 (web query 2004)

Table 2. Statistiques sur les collections et les requêtes utilisées

L'indexation, la recherche et l'évaluation sont réalisées en utilisant la plate-forme Terrier (Macdonald *et al.,* 2008), où les termes vides sont éliminés et l'algorithme de porter est utilisé en indexation et en recherche.

4.2 Evaluation

Nous avons utilisé les modèles suivants dans nos expérimentations :

ULM : Dans ce modèle la probabilité a priori de pertinence est ignorée. Le modèle de document est un modèle uni-gramme basé sur le lissage de Dirichlet.

ULM_IN : Modèle uni-gramme basé sur le lissage de Dirichlet, utilisant le nombre de liens entrants de la page (IN) pour estimer la probabilité a priori de document.

ULM_APP1 : Modèle uni-gramme basé sur le lissage de Dirichlet, utilisant la formule (2) (première version) pour estimer la probabilité a priori de document.

ULM_APP2 : Modèle uni-gramme basé sur le lissage de Dirichlet, utilisant la formule (3) (seconde version) pour estimer la probabilité a priori de document.

ULM_APP3 : Modèle uni-gramme basé sur le lissage de Dirichlet, utilisant la formule (4) (troisième version) pour estimer la probabilité a priori de document.

L'estimation de la pertinence finale d'un document (score) est réalisée, pour tous les modèles intégrant la probabilité a priori comme suit : Initialement, on effectue un classement des documents en tenant compte uniquement de leurs contenus, ensuite on effectue un reclassement des mille (1000) premiers documents retournés et cela en utilisant les deux évidences : le contenu de document et sa probabilité a priori.

Afin d'évaluer notre modèle et de le comparer aux autres modèles nous utilisons la mesure MAP (Mean Average Precision). La table ci-dessous montre les résultats obtenus avec les différents modèles cités précédemment. pour vérifier la significativité des résultats obtenus, nous avons effectué le test de Student et nous avons joint « $^{+}$ » et « $^{++}$ » pour l'indice de performance dans la table des résultats lorsque le test passe respectivement 95% et 99%.

Modèle	MAP	Amélioration
ULM	0,2404	
ULM_IN	0,2709	12,69%$^{++}$
ULM_APP1	0,2775	15,43% $^{++}$
ULM_APP2	0,2727	13,44% $^{++}$
ULM_APP3	**0,2801**	**16,51 %** $^{++}$

Table 3. Résultats des différents modèles sur la collection .GOV.

A partir des résultats obtenus et présentés dans la table 3, nous tirons les remarques et conclusions suivantes :

Premièrement, on remarque que l'utilisation de nombre de liens entrants comme seconde évidence a amélioré significativement le résultat obtenu avec le modèle uni-gramme (ULM), qui considère uniquement le contenu dans le classement des documents, l'amélioration constatée est de l'ordre de +12,69%. Ceci indique qu'une page populaire (ayant un nombre de liens entrants important) est plus probable d'être pertinente qu'une page moins populaire.

Deuxièmement, l'introduction d'une caractéristique de site (nombre de liens entrants) contenant une page dans le calcul de la probabilité a priori de cette page, permet d'améliorer le résultat obtenu avec le modèle considérant uniquement la caractéristique de la page (nombre de liens entrants), et cela avec les différentes versions proposées. Ceci indique que le calcul de la probabilité a priori d'une page est mieux estimé lorsque les caractéristiques de son site sont prises en compte.

Troisièmement, on remarque que la troisième version de notre modèle donne le meilleur résultat (+16,51%) par rapport aux versions (1) et (2), dans lesquelles on a obtenu respectivement des améliorations de +15,43% et +13,44%. On peut donc en déduire qu'une page doit hériter des caractéristiques de son site et cela selon sa similarité sémantique avec son site. Autrement dit une page traitant de la même thématique (similaire) que son site doit hériter plus de caractéristiques de son site qu'une page traitant une thématique différente de son site.

Afin d'avoir une vision plus fine et détaillée des améliorations obtenues par notre modèle, on a effectué une analyse requête-par-requête. Le graphique ci-dessous présente cette analyse.

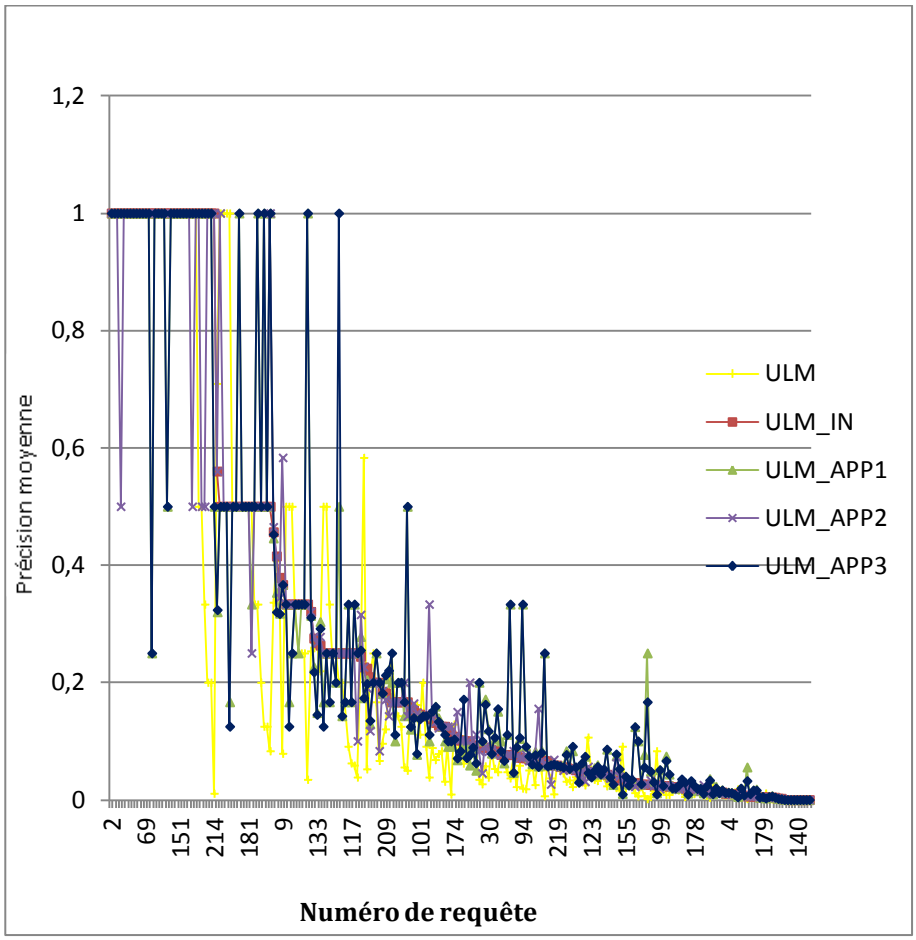

Figure 1. Résultats requête-par-requête avec les différents modèles de recherche.

À partir des résultats obtenus nous avons noté les remarques suivantes :

La première version de notre modèle améliore le modèle ULM_IN sur 86 requêtes ; ce dernier donne de meilleurs résultats sur 72 requêtes ; les deux modèles présentent des résultats équivalents sur 67 requêtes.

La seconde version de notre modèle améliore le modèle ULM_IN sur 57 requêtes ; ce dernier donne de meilleurs résultats sur 52 requêtes ; les deux modèles présentent des résultats équivalents sur 116 requêtes.

La troisième version de notre modèle améliore le modèle ULM_IN sur 87 requêtes ; ce dernier donne de meilleurs résultats sur 69 requêtes ; les deux modèles présentent des résultats équivalents sur 69 requêtes.

Dans l'optique de comprendre les améliorations constatées, nous avons analysé manuellement quelques requêtes. Ci-dessous, sont illustrés les détails des résultats de la requête numéro WT04-50 (money laundering).

Document	ULM	ULM_IN	ULM_APP1	ULM_APP2	ULM_APP3
G09-93-3919157	4	2	3	3	1
G10-04-2315842	14	9	6	7	6
G07-70-0000000	87	53	37	47	36
G35-71-1435292	117	109	106	111	102
G35-97-3029817	126	121	127	127	129
G22-50-2198391	252	287	305	304	300
G21-40-2003687	396	421	449	435	458
G08-69-0634583	458	580	692	677	728
G05-05-1028021	Non retrouvé	Non retrouvé	Non retrouvé	Non retrouvé	Non retrouvé
MAP	0,0622	0,1009	0,0969	0,1009	**0,1712**

Table 4. Le rang des documents pertinents avec les différents modèles

A partir de cette table on peut voir que la troisième version de notre modèle permet d'améliorer le rang des documents pertinents comparativement aux modèles ULM et ULM_IN. Par exemple le document *G09-93-3919157* est passé du rang 4 avec le modèle ULM à la 2ème place avec le modèle ULM_IN et à la 1ère place avec notre modèle ULM_APP3. Cette amélioration est expliquée principalement par le fait que le document *G09-93-3919157* contient 4 liens entrants et que son site contient 589 liens entrants et que ce document est similaire avec son site (0,0024 : valeur normalisée).

5. Conclusion

Nous avons proposé dans ce papier une approche permettant d'intégrer l'importance d'un site web dans le calcul de la probabilité a priori de pertinence d'une page web, et cela en se basant sur le postulat suivant : « on assigne plus de confidence aux pages provenant des sites importants (intéressants) que pour celles provenant des sites moins importants ».

Les expérimentations effectuées sur la collection .GOV ont montré que la prise en compte de l'importance d'un site web dans l'évaluation de la probabilité a priori de pertinence d'un site web améliore significativement le modèle uni-gramme, de plus les différentes versions de notre modèle affichent des améliorations par rapport au modèle considérant uniquement l'importance d'une page dans l'évaluation de sa probabilité a priori. Nous avons constaté également que la troisième version de notre

modèle affiche les meilleurs résultats en permettant à une page dont le contenu informationnel est similaire à celui de son site, de bénéficier plus de l'importance de site, plus que les autres pages.

Plusieurs pistes peuvent être investies dans le futur. En premier lieu, le test de nos approches sur d'autres collections web. Deuxièmement, l'usage d'autres facteurs de popularité tel que le PageRank. Enfin, propager la popularité des pages web vers leur site. En d'autres termes une page populaire renforcera la popularité de site qui la contient.

6. Bibliographie

Brin, S., Page, L., «The anatomy of a large-scale hypertextual web search engine». *In Proceedings of WWW7 (Brisbane, Australia, May 1998). http://www7.scu.edu.au/programme/fullpapers/1921/com1921.html*

Cohn, D., Chang, H., «Learning to probabilistically identify authoritative documents». *In Proceedings of the 17th International Conference on Machine Learning,* 2000.

Craswell, N., Robertson S., Zaragoza, H. and Taylor, M. «Relevance weighting for query independent evidence», In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2005, p. 416–423.

Diaz, F., Jones, R. «Using temporal profiles of queries for precision prediction». *In Proceedings of the 27th annual international conference on Research and development in information retrieval*, 2004, p. 18–24.

Hauff, C., Azzopardi, L. «Age dependent document priors in link structure analysis». *In The 27th European Conference in Information Retrieval*, 2005, p. 552–554.

Kraaij, W., Westerveld, T., and Hiemstra, D. «The importance of prior probabilities for entry page search», *In ACM International Conference on Research and Development in Information Retrieval*, 2002, p. 27–34.

Li, X., Croft, W.B. «Time-based language models». *In Proceedings of the twelfth international conference on Information and knowledge management*, 2003, p. 469–475.

Macdonald, C., and He, B., «Researching and Building IR applications using Terrier», *In European Conference on Information Retrieval*, 2008.

Parapar, J., E.Losada, D., Barreiro, A. «Compression-Based Document Length Prior for Language Model». *In ACM International Conference on Research and Development in Information Retrieval*, 2009.

Peng, J., Ounis I. «Combination of document priors in web information retrieval», In *Proceedings of European conference on information Retrieval*, 2007, p. 732–736.

Ponte, J. and Croft, W.B., «A language modeling approach to information retrieval», *In ACM International Conference on Research and Development in Information Retrieval,* 1998, p. 275-281.

Upstill, T., Craswell, N., Hawking, D. «Predicting fame and fortune: PageRank or indegree? », *In Proceedings of the Australasian Document Computing Symposium*, 2003.

Zhai, C., Lafferty, J., «A study of smoothing methods for language models applied to information retrieval», *ACM Transactions on Information Systems (TOIS)* Volume 22, Issue 2, 2004, p. 179 – 214.

Zhu, X.L., Gauch, S. «Incorporating quality metrics in centralized / distributed information retrieval on the world wide web», *In Proceedings of the 23th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2000, p. 288–295.